基于 SVM 多特征融合的微博情感多级分类研究*

杨　爽　陈　芬
(南京理工大学经济管理学院　南京 210094)

摘要:【目的】为更精确地识别网民态度, 监测网络舆情, 提出一种基于 SVM 多特征融合的情感 5 级分类方法。【方法】从词性特征、情感特征、句式特征、语义特征 4 个方面, 提取动词、名词、情感词、否定词等 14 个特征, 运用 SVM 方法对微博情感进行 5 级分类。【结果】实验结果表明, 该方法对情感 5 级分类的准确率为 82.40%, 召回率为 81.91%, F 值为 82.10%。【局限】训练语料的规模有待进一步提高。【结论】该方法在情感 5 级分类方面取得较好的效果。

关键词: 微博　情感倾向性　支持向量机　句法分析
分类号: G35　TP391

1　引　言

微博已成为中国当前用户数量最多的互联网信息传播平台。在微博中潜藏着极为丰富的主观情感信息。通过对微博进行情感分类, 获取广大网民们的情感倾向, 可以迅速、准确地了解广大网民的诉求, 为网络舆情分析提供可靠依据。

目前, 已有许多学者对微博情感分类进行研究, 主要采用基于语义的方法和基于机器学习的方法, 将情感分为正面、负面, 或者正面、中性和负面。然而这种划分方法并不能精确地反映网民们的情感立场[1]。在网络舆情中, 部分网民会表达自己对某事件的绝对立场, 他们很难受其他言论的影响。而有的网民表现的立场并不稳定, 他们只是暂时性受到某些言论的影响, 表现出倾向性的立场。所以, 将情感划分为三级过于绝对化, 应该采用非常正面、正面、中立、负面、非常负面的 5 级分类方法。而在现有的情感分类研究中, 大多是对以产品评论为主的中长文本进行 5 级分类, 对微博短文本的情感 5 级分类研究较少。

本文采用 SVM(Support Vector Machine)模型作为分类模型, 从词性特征、情感特征、句式特征和语义特征 4 个方面考虑, 提取词性、情感词、情感强度、情感得分和语义关系等多种情感资源特征, 对微博进行 5 级情感分类。

2　相关研究

目前文本情感分类技术主要有两类, 基于情感词典的方法和机器学习的方法。基于情感词典的方法是通过构建情感词典, 通过特定的算法模型进行情感倾向值的计算, 进而根据情感倾向值对文本进行极性分析。Kamps 等[2]利用 WordNet 的同义结构图计算新词与种子词的语义距离, 并以此计算情感倾向。Shen 等[3]构建了否定词词典、程度副词词典、感叹词词典和情感词词典, 设置相应的规则计算微博的情感倾向性, 准确率达到 80.6%。郑诚等[4]以情感词典的构建为基础, 将情感词、否定词、程度副词之间的语义规则加

通讯作者: 杨爽, ORCID: 0000-0003-3101-3544, E-mail: doubleyou1001@163.com。

*本文系国家自然科学基金项目“基于情感倾向性分析的网络舆情意见领袖识别与对策研究”(项目编号: 71303111)、国家自然科学基金项目“突发事件网络舆情演变过程中的人群仿真研究”(项目编号: 71273132)和国家自然科学基金项目“基于聚合的社会化短文本信息处理与细粒度倾向性分析”(项目编号: 71503126)的研究成果之一。

入微博的情感倾向计算中，之后结合情感词典与规则，计算微博的情感极性值，实现微博情感分类。机器学习的方法是将情感分类看作一种特殊的文本分类，通过机器学习算法训练标注好的训练集得到分类模型，再由分类模型确定文本的倾向性[5]。Pang 等[6]将机器学习方法应用到文本情感分类的研究中，发现选取 Unigram 为特征并结合 SVM 算法时的效果最好，分类正确率最高达到约 80%。Barbosa 等[7]使用从三个不同的 Twitter 情感分析网站上获取到的训练数据训练标准的 SVM 分类器，精度达到 81.3%。Davidov 等[8]使用微博中的标签、表情符号等作为特征，训练了一个类似 KNN 的分类器，将情感分为正、负两类，正确率最高达到 86%。夏梦南等[9]在进行微博的情感分析时，利用句法分析和 CRFs 抽取候选评价对象，以此为基础使用 SVM 方法对微博进行情感分类，正确率达到 91.4%。

通过以上的研究分析发现，目前的情感分类研究仍多以三级分类为主，并且已经有较高的准确率。而在现实应用中，三级分类并不能很好地满足实际需求，尤其是在产品评论方面，所以不少学者对文本的 5 级分类进行研究。Ding 等[10]通过改进条件随机场(CRFs)，分两个层次两次使用 CRFs 方法。第一层对文章进行极性判断，第二层给出 5 个级别的强度分类，取得了相对不错的效果。魏晶晶等[11]对电子商务产品评论进行多级情感分析，评论最终被划分为：强烈贬意、一般贬意、中性、一般褒扬、强烈褒扬 5 级的情感强度，通过复杂句句法模式和基于词典的算法计算句子级情感倾向，进而得出整个评论的分类。该方法主要针对篇章级文本进行 5 级分类，对于句子级 5 级分类并没有深入研究。廖健等[12]提出一种基于观点袋模型和语言学规则的多级情感分类方法。该方法通过计算搭配四元组的情感倾向极性值，建立文本的向量化表示，并构造权重计算公式，利用文本余弦相似度计算方法实现对汽车评论文本的 5 级情感极性分类。但该方法需要使用已有的领域本体特征，抽取的情感搭配无法覆盖全部文档。上述方法主要是针对产品评论进行多级分类，而微博相比于产品评论，文本长度更短，表达更随意，目前对于微博短文本的多级情感分类研究还较少。

本文在已有研究的基础上，通过 Word2Vec 发现网络情感新词，并加入语义特征，通过句法依存技术，获取句子中与情感词有关的语义关系，提出一种融合多种情感资源特征，利用 SVM 分类器实现微博情感 5 级分类的方法。

3 基于 SVM 多特征融合的微博情感 5 级分类

3.1 词典构建

根据情感分析的需要，构建了三个词典：情感词典、否定词词典、程度副词词典。本文以知网 HowNet 情感词词典为基础，并利用 Word2Vec[13]发现网络情感新词。Word2Vec 是利用词语间的语义关系，将词语转化为词向量，然后利用词向量之间的语义距离关系，自动识别网络情感新词。其原理是用哈夫曼树构建生成的统计语言模型中的概率模型，针对训练语料，利用浅层神经网络后向传播算法(Back Propagation，BP)传递误差损失，同时更新神经网络中的模型参数与词向量，通过若干轮的迭代生成该统计语言模型，并同时生成语料中所有词汇的词向量，如公式(1)所示。

$$\theta, C = \arg\max_{\theta, C} \frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leqslant j\leqslant c, j\neq 0} \log p(w_{t+j} \mid w_t; \theta, C) \tag{1}$$

其中，θ表示模型中神经网络的相关参数，C 表示语料所有词汇所构成的矩阵向量 $V\times K$(V 表示词汇个数，K 表示词向量纬度)。若使用哈夫曼树的数据结构，公式(1)中的 $p(w_{t+j} \mid w_t; \theta, C)$ 定义如公式(2)所示。

$$p(w_{t+j} \mid w_t; \theta, C) = \prod_{i=2}^{l^{w_y}} p(d_i^{w_y}; C_{w_x}, \theta_{i-1}^{w_y})$$
$$= \prod_{i=2}^{l^{w_y}} [\sigma(C_{w_x}^T \cdot \theta_{i-1}^{y})]^{1-d_i^{w_y}} \cdot 1-[\sigma(C_{w_x}^T \cdot \theta_{i-1}^{w_y})]^{d_i^{w_y}} \tag{2}$$

其中，l^{w_y} 表示从根节点出发到 w_y 所对应的叶子节点中所包含的非叶子节点个数，这些节点相对应的哈夫曼编码分别为 $d_i^{w_y}$，对于神经网络中权重参数为 $\theta_{i-1}^{w_y}$，C_{w_x} 表示 w_x 的词向量，而σ(x)是 sigmod 函数，定义如公式(3)所示。

$$\sigma(x) = \frac{1}{1+e^{-x}} \tag{3}$$

通过窗口的滑动，当模型完成对语料的数次迭代学习后，得到其统计语言模型相关参数θ与所有词汇组成的词向量矩阵 C。

通过使用 Word2Vec 对情感词语进行扩充，并进行人工筛选和调整。最终，情感词典包括 4 566 个正向情感词和 4 371 个负向情感词。否定词以《中国现代语法》中给出的否定词为基础，并对否定词词典进一步扩展，最终得到 28 个否定词。程度副词以 HowNet 情感词典中的程度副词词典为基础，再通过人工收集，最终得到 256 个程度副词。本文对不同语气强度的程度副词，分别赋予 0.5 到 2 的权重。部分程度副词及其权重如表 1 所示。

表 1 程度副词表

权重	示例	个数
2.0	百分之百、绝对、非常、超、过于……	99
1.5	很、多么、更加、不胜……	78
1.0	比较、较为、多多少少……	13
0.5	稍微、略为、不怎么、不为过……	54

3.2 特征选择

不同级别的微博在语义及语法结构上均存在不同特征。特征选择是使用 SVM 分类器分类的重要环节[14]，分类结果的准确率、召回率及分类系统的效率均取决于特征选择的合理性。本文通过阅读文献以及观察真实微博语料，从词性特征、情感特征、句式特征和语义特征 4 个方面，提取 13 个特征，如表 2 所示。

表 2 特征类型及含义

特征类型	含义
词性特征	微博中含有的动词数量(F1)
	微博中含有的形容词数量(F2)
	微博中含有的副词数量(F3)
情感特征	微博中含有的正面情感词数量(F4)
	微博中含有的负向情感词数量(F5)
	微博中程度副词的最高权重(F6)
	微博的情感得分(F7)
句式特征	否定词的数量(F8)
	感叹号的数量(F9)
	问号的数量(F10)
语义特征	与情感词有关的副词性修饰语(F11)
	与情感词有关的形容词性修饰语(F12)
	与情感词有关的名词性主语(F13)

(1) 词性特征

微博语言的特点就是短小、精悍。用户在使用微博发布自己的观点、想法时，有时仅仅采用一个或多个词语表达自己的想法，并没有完整的结构。所以将词性考虑到情感倾向的特征内，可以更好地解析句子的结构，辅助模型判断微博的情感。根据文献[15]以及对语料的观察，本文选择动词、形容词、副词作为情感分类的特征。

(2) 情感特征

情感词是最能直观反映微博发布者情感状态的词语。情感词一般分为“正面情感词”、“负面情感词”。正面情感词是指词语本身表现出比较积极、乐观等态度；负面情感词是指词语本身表现出失落、消极等态度。将正面情感词和负面情感词作为情感分类的特征。

要实现对情感的 5 级分类，情感强度显得尤为重要。在本研究中，情感强度是通过情感词前出现的程度副词的权重体现。比如：“她长得非常好看”，正面情感词“好看”之前出现程度副词“非常”，“非常”的权重为 2，所以“好看”的情感强度由 1 变成 2。对于一条微博中出现多个程度副词，取强度最高的值作为情感强度特征值。本文还将情感得分作为特征之一。情感得分高的句子比情感得分低的句子情感倾向更明显。情感得分计算如公式(4)所示。

$$Score = \sum_{i=0}^{n} (rawscore_i \times Intense_i) \tag{4}$$

其中，n 是一条微博中的句子数，$rawscore_i$ 是第 i 个句子中情感词的基本分数(+1、–1 或 0)；$Intense_i$ 是第 i 个句子的修饰词程度权重或否定词权重。

(3) 句式特征

否定词的出现可能会改变整个语句的情感倾向。如“今天玩的不开心!”，如果不考虑否定词，该文本的情感倾向性归类为正面，但是该句真实的情感倾向为负面。由此可知，否定词是情感倾向分析过程中比较重要而且是必不可少的特征。

如果一句话中出现问号和感叹号，说明微博发布者在强调自己的情感，问号和感叹号出现的次数不同，所表达情感程度也不同。感叹号和问号作为微博情感倾向的特征会更好地辅助模型判别微博的情感倾向性。因此将问号和感叹号出现的次数作为情感分类的特征之一。

(4) 语义特征

句法分析是指对输入的单词序列(一般为句子)判断其构成是否为合乎给定的语法，分析出合乎语法的

句法结构[16]。通过依存句法分析，能够体现微博的内部结构和联系，更能全面地表现微博的情感倾向。本文采用 Stanford Parser[17]句法分析器进行句法分析。通过对真实语料的观察，并参考文献[18]，提取以下三种关系类型作为特征。

①advmod 副词性修饰语：副词性修饰语用于改变该副词的强度。例如，"她长得非常漂亮"的提取结果是 advmod(漂亮，非常)，表示"非常"作为副词修饰了"漂亮"这个形容词。

②amod 形容词修饰语：一个名词词组的形容词修饰语。例如"这可真是神回复啊"的提取结果是：amod(回复，神)，表示名词性形容词"神"修饰了"回复"。

③nsubj 名词性主语：用于修饰名词性主语。例如"不一样的抗日神剧，好看！"的提取结果是 nsubj(好看，剧)，表示"好看"修饰了名词性主语"剧"。

3.3 情感分类模型

微博语料含有#话题#、URL 和@用户等无用信息，这些信息并不包含用户的观点，还可能影响下一步分词和词性标注的效果。因此在分词之前，首先过滤掉微博中的#话题#、URL 和@用户等无用信息，然后再对过滤后的语料进行下一步处理。本文使用中国科学院计算技术研究所的分词工具 NLPIR2016[19]对过滤后的语料进行分词和词性标注。选择 SVM 模型作为情感分类模型。实验使用所选择和计算得到的特征来表示训练集和测试集，之后将训练集输入 SVM 模型中，构建微博情感分类模型并对模型参数进行优化，最后将测试集输入构建好的模型中，得到测试集的情感类别。情感分类模型如图 1 所示。

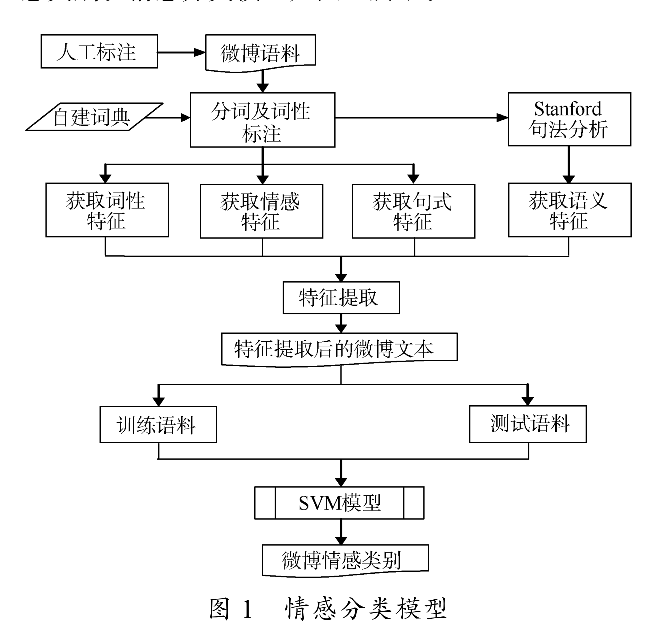


图 1 情感分类模型

4 实验设计

4.1 实验语料及标注标准

实验数据使用部分 COAE2014 微博评测语料，人工对这些语料按照"非常正面"、"正面"、"中立"、"负面"、"非常负面"5 个情感级别进行标注。标注工作由课题组成员完成，共标注 5 000 条语料。标注结果如表 3 所示。

表 3 实验数据分布

类别	数量
非常正面	217
正面	1 149
中立	2 081
负面	1 239
非常负面	304

人工标注语句的情感值主要依据语句中含有的程度副词级别和标点符号判断。语句中如果含有如"非常"等这样程度副词级别较高的词语，这样的语句比程度副词级别较低或不含程度副词的语句情感表达更强烈。同样地，如果语句中含有多个"!"或"?"等具有表达情感的标点符号，这样的语句也比不含任何标点符号的语句情感表达更强烈。

例句(1)：这个翡翠挺好看！

例句(2)：这个翡翠真的非常非常好看!!!

例句(2)比例句(1)的程度要更强烈，所以例句(1)的情感值标注为+1，而例句(2)的情感值标为+2。

4.2 特征提取结果

完成标注后，对实验语料进行分词、词性标注等操作，并按 3.2 节所述特征提取方法提取特征。实验使用 Java 语言编写程序，在 Eclipse 平台下完成所有程序编写及测试。实验环境是 Win7 64bit 操作系统，内存 4GB。表 4 为部分文本内容特征抽取结果。

4.3 模型及评价指标

本文使用 LibSVM 进行 SVM 模型的训练与分类[20]。将每类情感语料按 4：1 分成训练集和测试集。在训练语料前，对提取的特征进行归一化处理，以提高运行速度。训练采用 LibSVM 提供的默认参数，SVM 类型为 C_SVC，核函数为 RBF 核函数。采用准确率、召回率和 F1 值作为评价标准。

表 4　部分特征提取结果

特征 / 情感值	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13
+2	1: 2	2: 0	3: 2	4: 2	5: 0	6: 2.0	7: 4.0	8: 0	9: 3	10: 0	11: 1	12: 0	13: 1
+1	1: 4	2: 2	3: 3	4: 3	5: 0	6: 0.0	7: 1.0	8: 0	9: 1	10: 0	11: 0	12: 2	13: 2
–2	1: 2	2: 2	3: 0	4: 0	5: 2	6: 2.0	7: -4.0	8: 1	9: 0	10: 1	11: 2	12: 0	13: 0
–1	1: 3	2: 5	3: 3	4: 1	5: 4	6: 1.0	7: -2.0	8: 3	9: 0	10: 6	11: 1	12: 3	13: 0
0	1: 3	2: 2	3: 3	4: 1	5: 0	6: 0	7: 1.0	8: 2	9: 1	10: 1	11: 2	12: 3	13: 0

5　实验结果及分析

(1) 不同特征组合对情感分类影响

通过实验验证不同情感特征构建方式对分类的影响，使用准确率作为评估指标，实验结果如表 5 所示。

表 5　不同特征组合实验结果

实验	特征组合	准确率
1	词性	57.60%
2	词性+情感词	80.93%
3	词性+情感词+程度副词权重	81.76%
4	词性+情感词+程度副词权重+情感得分	81.95%
5	词性+情感词+程度副词权重+情感得分+否定词	82.14%
6	词性+情感词+程度副词权重+情感得分+否定词+问号和感叹号	82.22%
7	词性+情感词+程度副词权重+情感得分+否定词+问号和感叹号+语义特征	82.40%

由表 5 可以看出，当采用所有特征时，准确率最高，达到 82.40%。其中情感词特征的作用最大，准确率提高了 23.33%，其次为程度副词权重，使准确率提高了 0.83%，其余特征也均对情感分类起到一定作用，使准确率有略微提升。

(2) 对比实验评价

将本文方法与文献[10]提出的层叠 CRFs 方法进行对比。层叠条件随机场模型(Cascaded CRFs, CCRFs)是按层叠加建立起多个层次的条件随机场模型，在 5 级分类中较为常用。通过该方法可以将 5 分类问题转为由粗到细的过程，通过低层模型识别出初步结果，进行过滤和整合，将处理后的识别结果输入到高层，为高层条件随机场提供决策支持。文献[10]采用层叠 CRFs 模型，首先对文本进行三级分类，然后结合词性特征、评价词特征、连词特征以及极性特征(即三级分类的结果)对情感进行 5 级分类，在 COAE2008 的任务 3 上，取得了很好的分类效果，准确率最高达到 83.75%。使用该方法在本文语料集上进行实验，并与本文方法进行对比，结果如表 6 所示。

表 6　对比实验结果

方法	准确率	召回率	F1 值
本文方法	82.40%	81.91%	82.10%
层叠 CRFs 方法	75.31%	73.30%	74.30%

由表 6 可以看出，本文提出的方法在 5 级分类的准确率为 82.40%，相较于层叠 CRFs(75.31%)，准确率有较大的提高。召回率为 81.91%，相较于层叠 CRFs(73.30%)，也有较大的提升。F 值综合考虑了准确率和召回率，本文方法的 F1 值为 82.10%，与层叠 CRFs(74.30%)相比，提升了 7.80%。文献[10]的层叠 CRFs 方法所提取的特征主要针对中长文本，对于微博短文本并不适用，所以准确率有所下降。本文利用 Word2Vec 对情感词典进行扩充，并从词性特征、情感特征、句式特征和语义特征多个维度对特征进行选择，在对微博进行情感 5 级分类中取得较高的准确率。

6　结　语

本文提出一种采用 SVM 分类器对微博语句进行 5 级情感分类的方法。该方法以词性特征、情感特征、句式特征、语义特征等作为依据，对微博语句进行 5 级情感分类，并将该方法与已有的 5 级分类方法进行对比，取得较好的效果。

本文的不足之处在于：训练语料较少，尤其是非常正面和非常负面这两类语料数量太少。一般来说，训练语料越多，构建的模型越准确，未来研究要扩大训练语料，进一步提高模型的准确性。

参考文献：

[1] 王雪猛，王玉平．基于情感倾向分析的突发事件网络舆情预警研究[J]．西南科技大学学报：哲学社会科学版，2016,

33(1): 63-66. (Wang Xuemeng, Wang Yuping. Research of Emergency Network Public Sentiment Warning Based on the Analysis of Emotional Tendency [J]. Journal of Southwest University of Science and Technology: Philosophy and Social Science Edition, 2016, 33(1): 63-66.)

[2] Kamps J, Marx M, Mokken R J, et al. Using WordNet to Measure Semantic Orientations of Adjectives [C]// Proceedings of the 4th International Conference on Language Resources and Evaluation. 2004.

[3] Shen Y, Li S, Zheng L, et al. Emotion Mining Research on Micro-blog [C]// Proceedings of the 1st IEEE Symposium on Web Society. 2009.

[4] 郑诚，杨希，张吉赓. 结合情感词典与规则的微博情感极性分类方法[J]. 电脑知识与技术，2014, 10(13): 3111-3113. (Zheng Cheng, Yang Xi, Zhang Jigeng. Micro-blog Sentiment Analysis of Combined Sentiment Dictionary and Rules [J]. Computer Knowledge and Technology, 2014, 10(13): 3111-3113.)

[5] 张阳，刘晓霞，孙凯龙，等. 基于情感描述项的文本倾向性识别研究[J]. 计算机工程与应用，2015, 51(4): 158-161, 195. (Zhang Yang, Liu Xiaoxia, Sun Kailong, et al. Research on Text Orientation Identification Based on Emotional Description Item [J]. Computer Engineering and Applications, 2015, 51(4): 158-161, 195.)

[6] Pang B, Lee L, Vaithyanathan S. Thumbs up? Sentiment Classification Using Machine Learning Techniques [C]// Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing. 2002.

[7] Borbosa L, Feng J. Robust Sentiment Detection on Twitter from Biased and Noisy Data [C]//Proceedings of the 23rd International Conference on Computational Linguistics. Beijing: Tsinghua University Press. 2010.

[8] Davidov D, Tsur O, Rappoport A. Enhanced Sentiment Learning Using Twitter Hashtags and Smileys [C]// Proceedings of the 23rd International Conference on Computational Linguistics: Posters, 2010: 241-249.

[9] 夏梦南，杜永萍，左本欣. 基于依存分析与特征组合的微博情感分析[J]. 山东大学学报：理学版，2014, 49(11): 22-30. (Xia Mengnan, Du Yongping, Zuo Benxin. Micro-blog Opinion Analysis Based on Syntactic Dependency and Feature Combination [J]. Journal of Shandong University: Natural Science, 2014, 49(11): 22-30.)

[10] Ding S, Jiang T, Wen N. Research on Sentiment Orientation of Product Reviews in Chinese Based on Cascaded CRFs Models [C]//Proceeding of the 2012 International Conference on Machine Learning and Cybernetics (ICMLC 2012). IEEE, 2012.

[11] 魏晶晶，吴晓吟. 电子商务产品评论多级情感分析的研究与实现[J]. 软件，2013, 34(9): 65-67, 94. (Wei Jingjing, Wu Xiaoyin. Research on Multi-level Sentiment Analysis System of E-Commerce Product Review and Implementation [J]. Software, 2013, 34(9): 65-67, 94.)

[12] 廖健，王素格，李德玉，等. 基于观点袋模型的汽车评论情感极性分类[J]. 中文信息学报，2015, 29(3): 113-120. (Liao Jian, Wang Suge, Li Deyu, et al. The Bag-of-Opinions Method for Car Review Sentiment Polarity Classification [J]. Journal of Chinese Information Processing, 2015, 29(3): 113-120.)

[13] Word2Vec [EB/OL]. [2015-01-12]. http://word2vec.googlecode.com/svn/trunk/.

[14] Liu Z, Yu W, Chen W, et al. Short Text Feature Selection for Micro-blog Mining [C]//Proceedings of the 2010 International Conference on Computational Intelligence and Software Engineering. IEEE, 2010.

[15] 吴明芬，陈涛. 基于SVM的以词性和依存关系为特征的句子倾向性判断分析[J]. 五邑大学学报：自然科学版，2012, 26(4): 66-71. (Wu Mingfen, Chen Tao. Sentences Tendency Judgement by POS and Dependency Based on SVM [J]. Journal of Wuyi University: Natural Science Edition, 2012: 26(4): 66-71.)

[16] 刘海涛. 依存语法的理论与实践[M]. 北京：科学出版社，2009. (Liu Haitao. Dependency Grammar: From Theory to Practice [M]. Beijing: Science Press, 2009.)

[17] Stanford Parser [EB/OL]. [2015-06-16]. http://nlp.stanford.edu/software/lex-parser.shtml.

[18] 彭玥. 基于文本倾向性的网络意见领袖识别[D]. 南京：南京理工大学，2014. (Peng Yue. Internet Opinion Leader Detection Based on Text Sentiment Analysis [D]. Nanjing: Nanjing University of Science and Technology, 2014.)

[19] NLPIR/ICTCLAS [EB/OL]. [2015-12-02]. http://ictclas.nlpir.org/.

[20] LibSVM [EB/OL]. [2015-07-12]. https://www.csie.ntu.edu.tw/~cjlin/libsvm/.

作者贡献声明:

杨爽：词典构建，程序设计，起草论文；

陈芬：提出研究思路，设计研究方案，论文最终版本修订。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据[1-3]见期刊网络版 http://www.infotech.ac.cn; 支撑数据[4]由作者自存储, E-mail: doubleyou1001@163.com。

[1] 杨爽, 陈芬. senti_dic.rar. 利用 Word2Vec 扩充后的正负情感词典.

[2] 杨爽, 陈芬. weight_dic.txt. 人工打分后, 含有权重的程度副词词典.

[3] 杨爽, 陈芬. TestData.rar. 人工标注的测试数据.

[4] 杨爽. Senti_analysis.rar. 特征选择程序.

收稿日期: 2016-08-29
收修改稿日期: 2016-10-26

Analyzing Sentiments of Micro-blog Posts Based on Support Vector Machine

Yang Shuang　Chen Fen
(School of Economics and Management, Nanjing University of Science & Technology, Nanjing 210094, China)

Abstract: [**Objective**] This paper proposes a new method based on the Support Vector Machine to monitor online public opinion. [**Methods**] We extracted fourteen linguistic characteristics of the micro-blog posts and analysed their sentiments with Support Vector Machine. [**Results**] The precision, recall and F value of the proposed method were 82.40%, 81.91%, and 82.10%, respectively. [**Limitations**] The size of training corpus needs to be expanded. [**Conclusions**] The proposed method could effectively analyze sentiments of micro-blog posts.

Keywords: Microblog　Sentiment Analysis　Support Vector Machine　Parsing